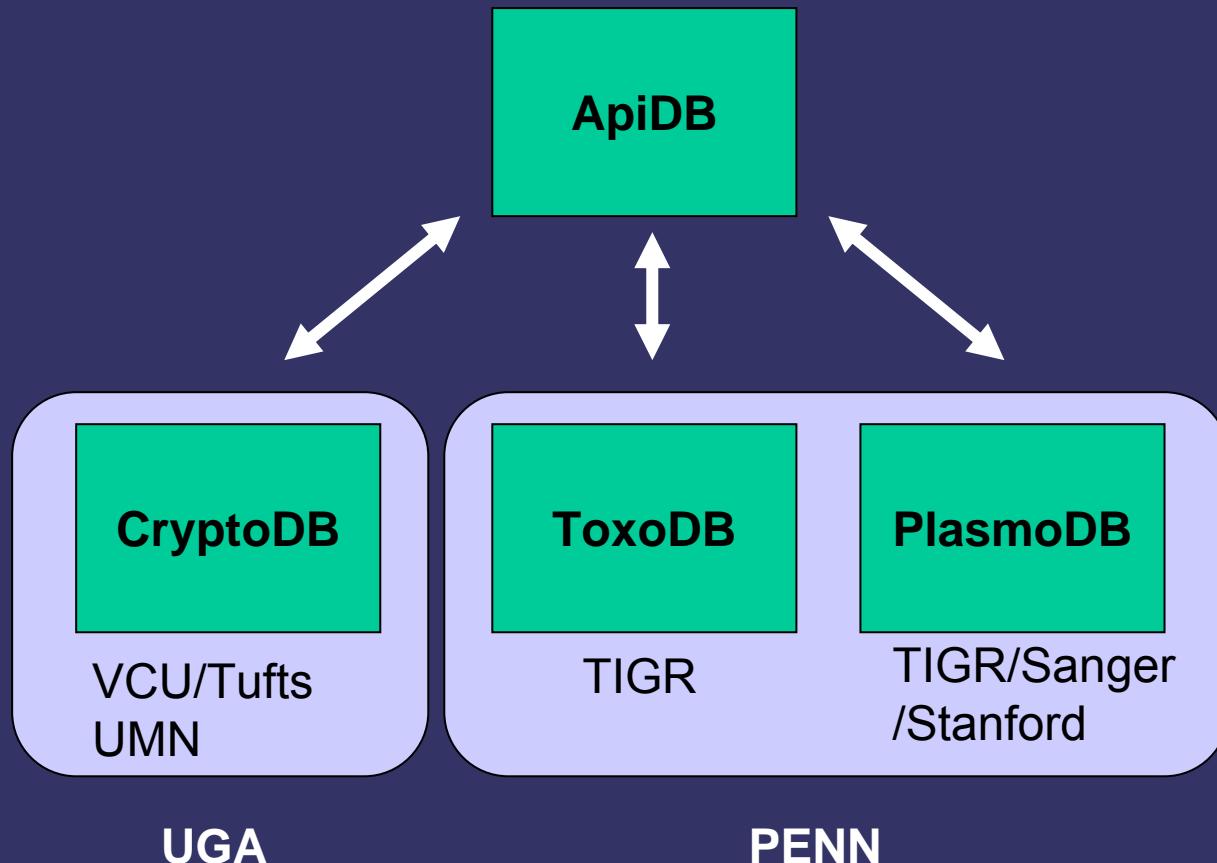


ApiDB: Our data types, curation status and analysis pipelines

Jessica Kissinger
ApiDB BRC

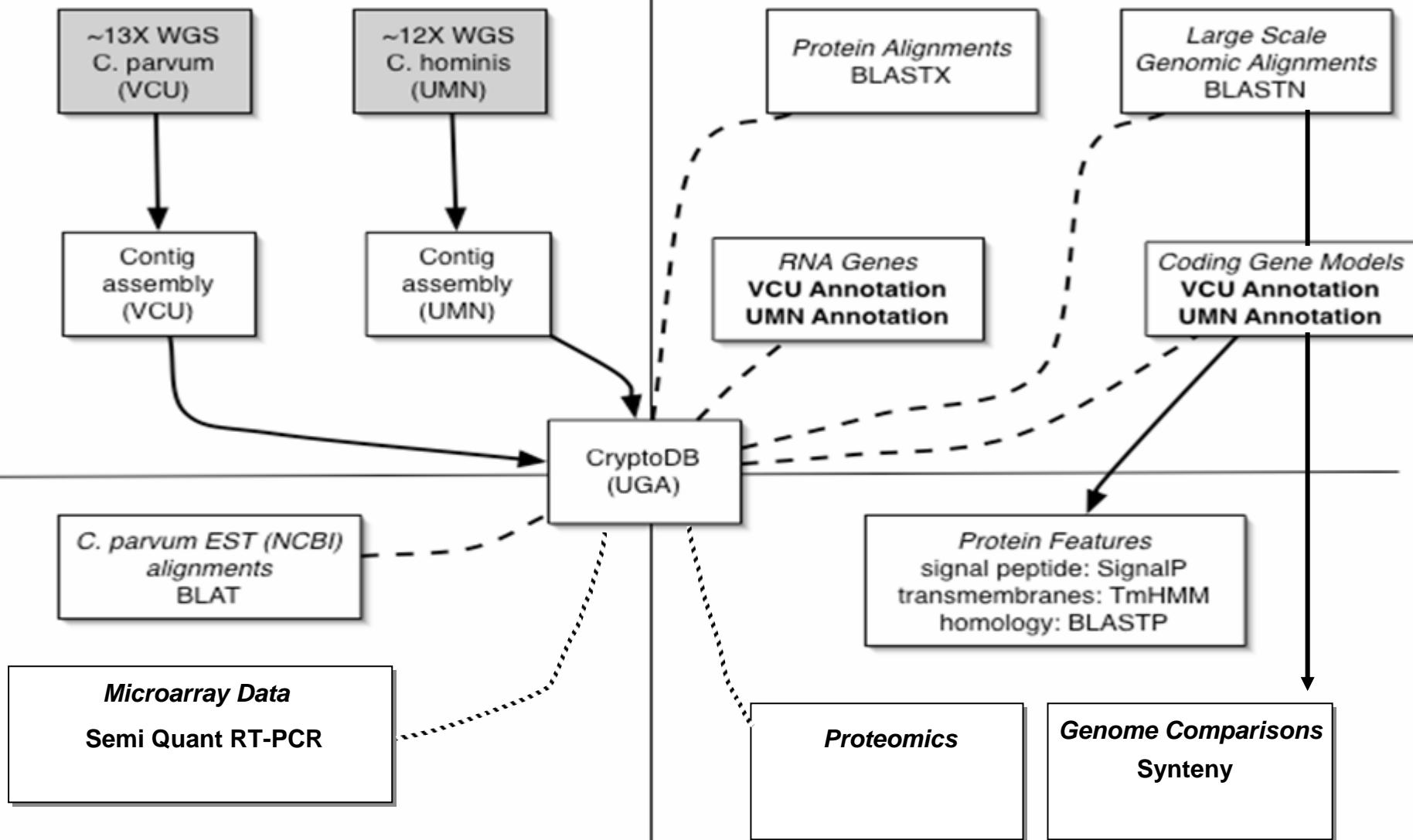
ApiDB

Structure and Genome sources

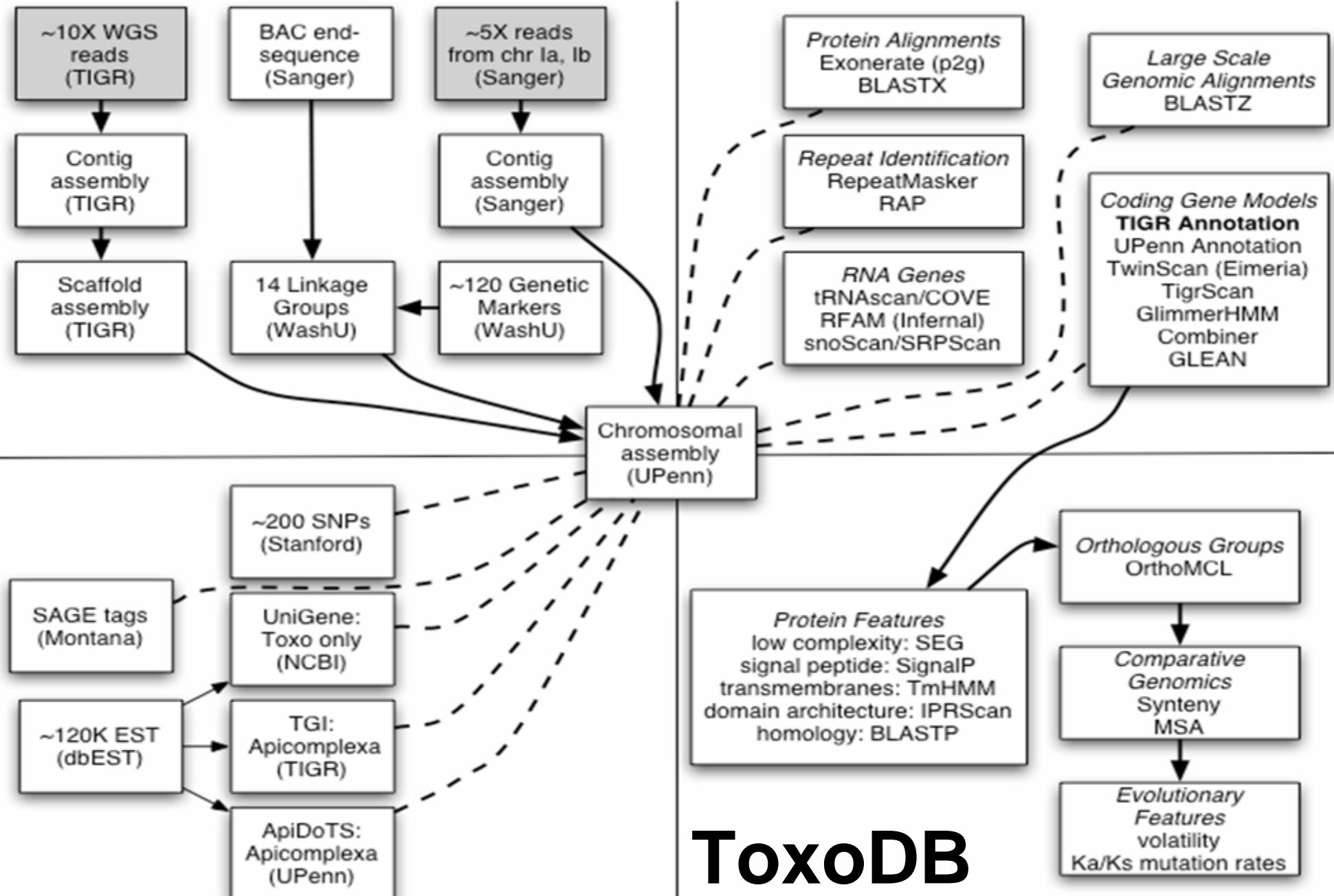


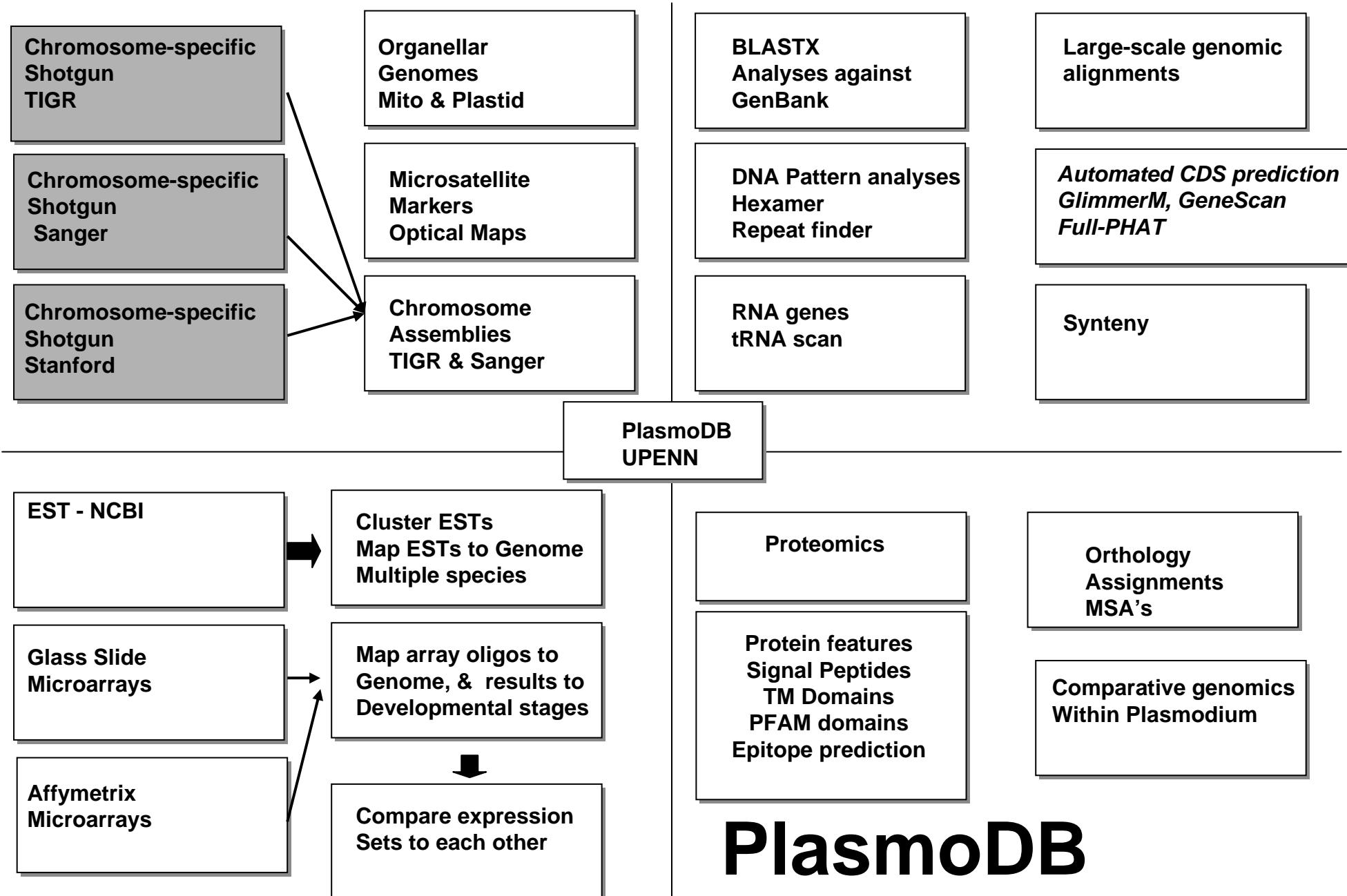
Annotation / Curation / Integration

	CryptoDB	ToxoDB	PlasmoDB
Genome(s)	2 (+4)	1 (+3)	5 (+3)
EST	567	125,741	69,600 (All Sp.) 20,914 (P.f.)
SAGEtags	X	>List icon	List icon
Microarrays	In Progress	In Progress	List icon
Proteomic	In Progress	In Progress	List icon
Microsatellite	X	List icon	List icon
SNP	?	List icon (ESTs)	In Progress
Genome Synteny	List icon	X	List icon
Orthologous genes	List icon	List icon	List icon
Metabolic Pathways	In Progress	In Progress	List icon
Curatorial control	No/Desired	No/Planned	No



CryptoDB





ApiComplexa Pipelines

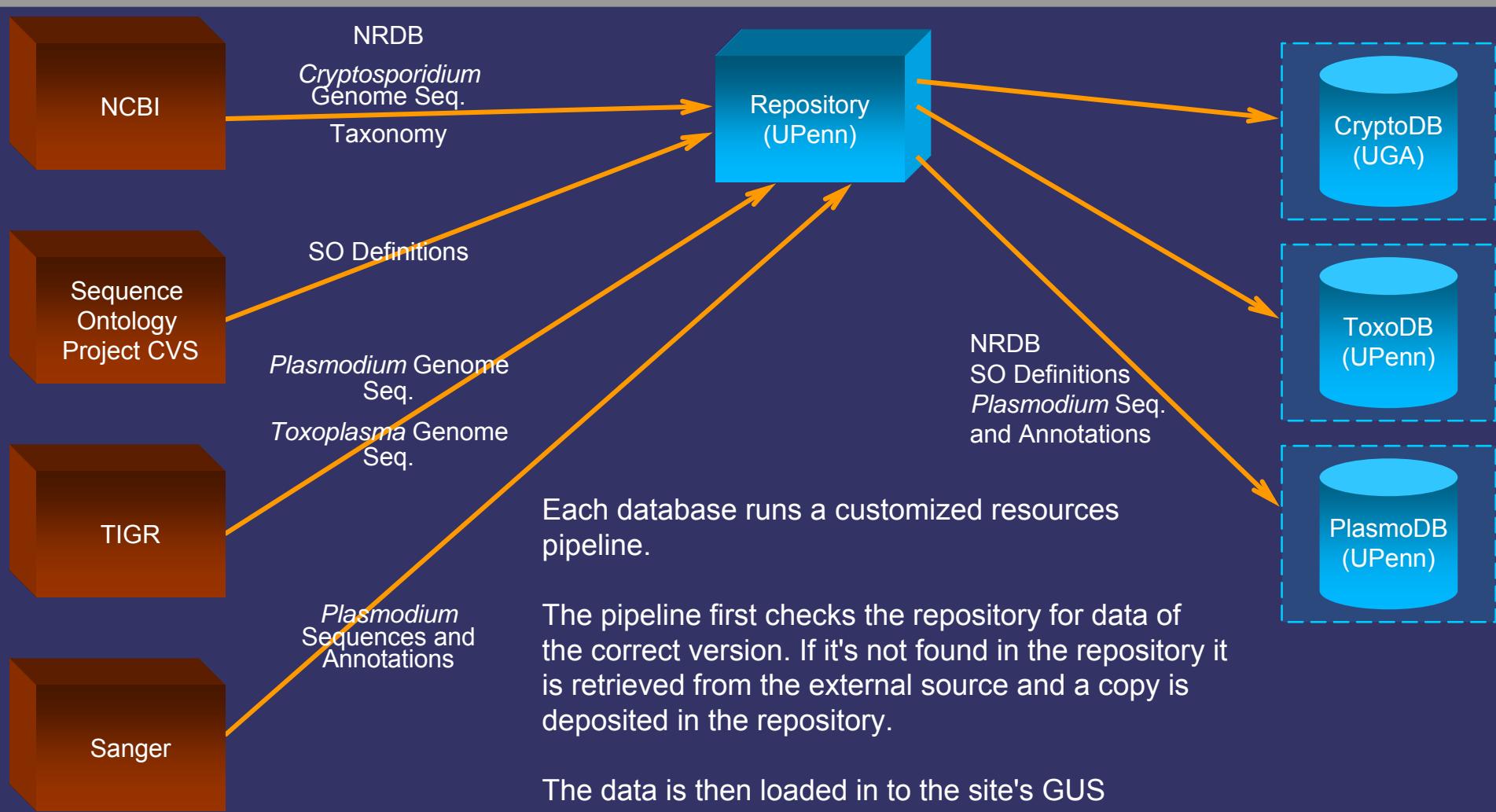
- External Resources Pipeline
 - Track all data entering the database(s)
- Data Analysis Pipeline
 - Provide automated, quality controlled analyses on primary data entered into the database
 - Provide additional features
 - Integrate with existing data
 - Integrate with external resources

Apicomplexa Pipelines

▷ Objectives

- Automation
 - acquire and archive data from disparate sources
 - load data into GUS database
 - perform analyses on raw data and load results in to database
 - Integrate diverse data sets to permit complex queries
- Standardization
 - ensures common data (nrdb, SO) are of the same version among individual sites
 - ensures common data analyses are performed the same way and with the same parameters on each site
- Organization
 - help manage the many data types: contigs, ESTs, annotations, SAGE tags, SNPs, genetic markers, etc.
 - help coordinate and log the myriad of data analyses performed
 - Provide very detailed release notes to our users

External Resources Pipeline



Resources Pipeline

XML Configuration

```
<resource resource="nrdb" version="2005-01-27"
          url="ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/"
          plugin="GUS::Common::Plugin::LoadNRDB"
          extDbName="NRDB"
          extDbRlsVer="continuous ApiDB"
          extDbRlsDescrip =">Latest release of this database"
          dbCommit=@dbcommit@>

<wgetArgs>--tries=5 --mirror --no-parent --no-directories
           --no-host-directories --cut-dirs=3 --accept=nr.gz
</wgetArgs>

<unpack> gunzip @downloadDir@/nrdb/nr.gz </unpack>

<pluginArgs>--temp_login @nrdb.tempTableLogin@
           --temp_password @nrdb.tempTablePassword@
           --sourceDB @nrdb.sourceDB@
           --dbi_str @nrdb.tempTableDbi@
           --restart @nrdb.restart@
</pluginArgs>

</resource>
```

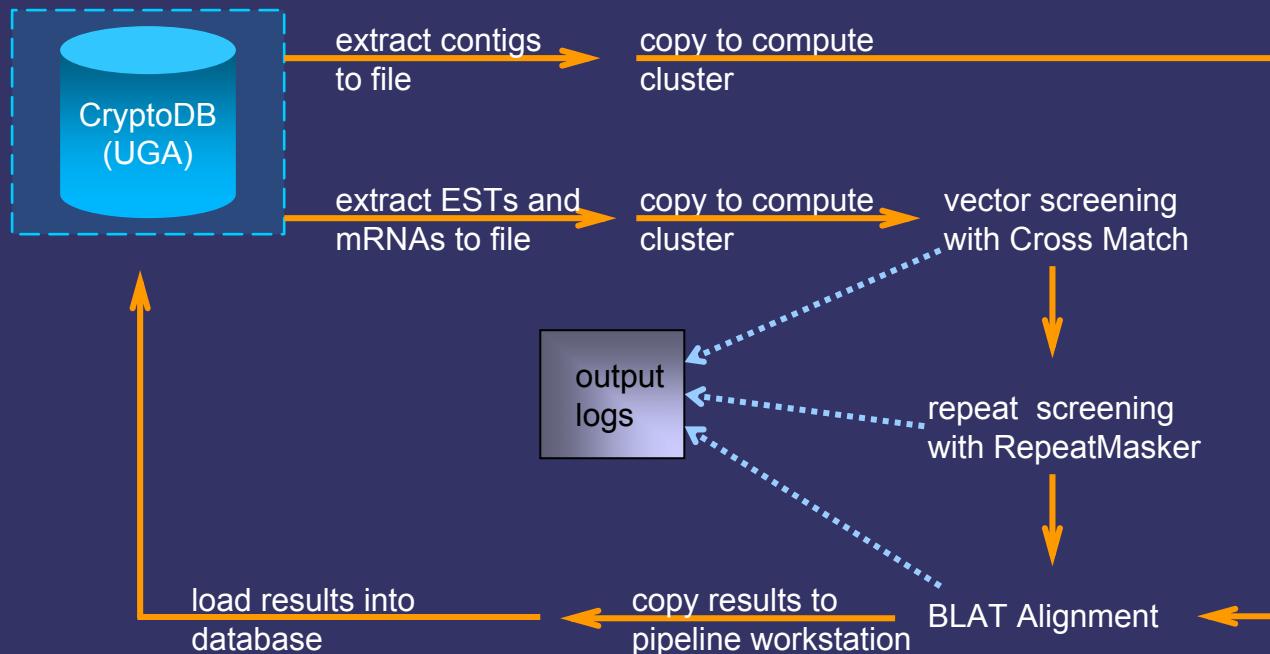
Analysis Pipeline

Library of ‘Steps’ - Perl subroutines and modules that perform data preparation (eg. extract data from GUS in appropriate format for analysis), wrap analysis programs (eg. BLAST, SignalP, Glimmer), and load analysis results into GUS

Steps from the library are chained together to perform the desired tasks for a given site.

Sample Step

Align ESTs to Contigs



ApiComplexa DAS Service

The Distributed Annotation System (DAS)

Towards A Comprehensive Community of
Annotations for ApiComplexan Genomes

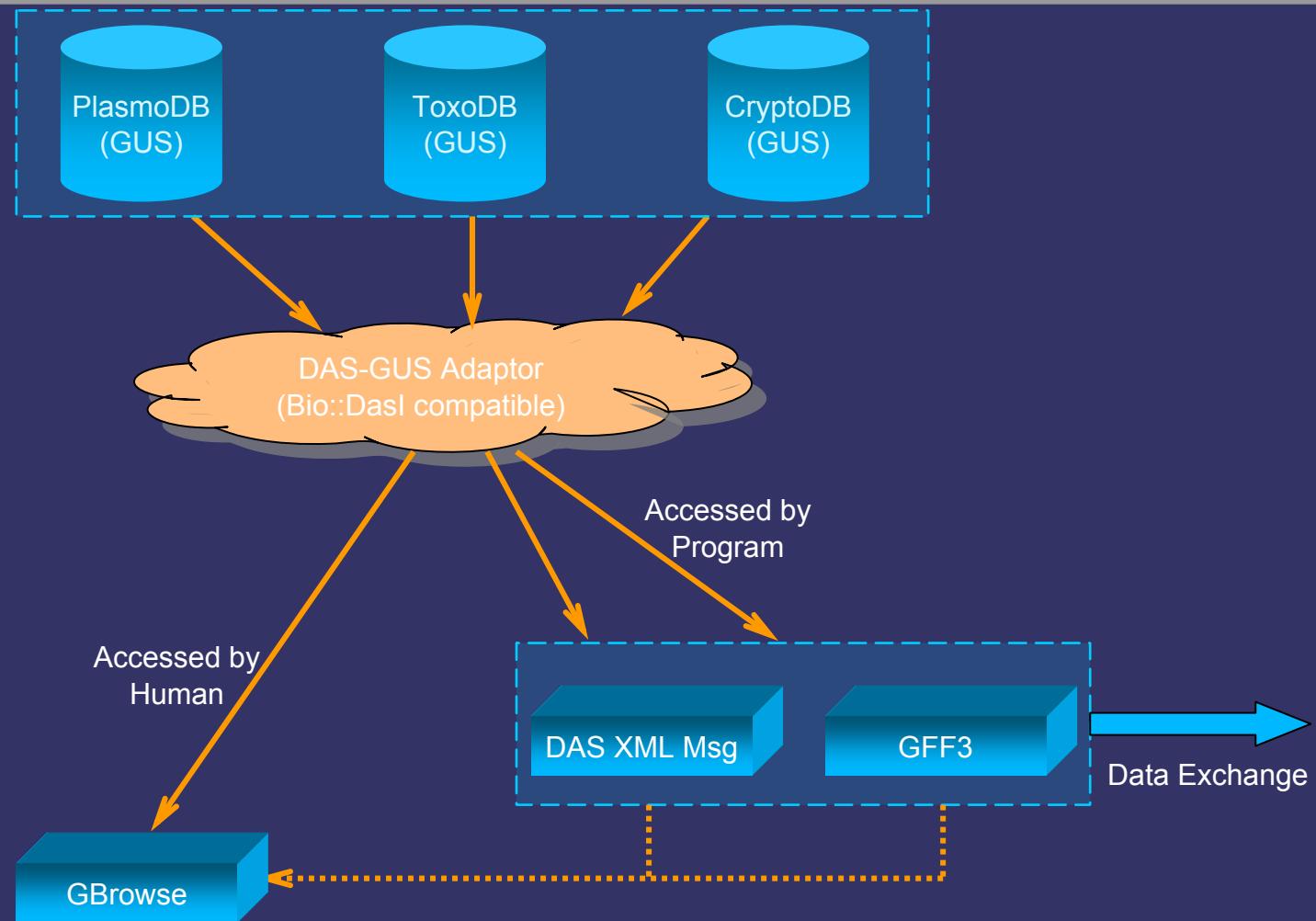
ApiComplexa DAS Service

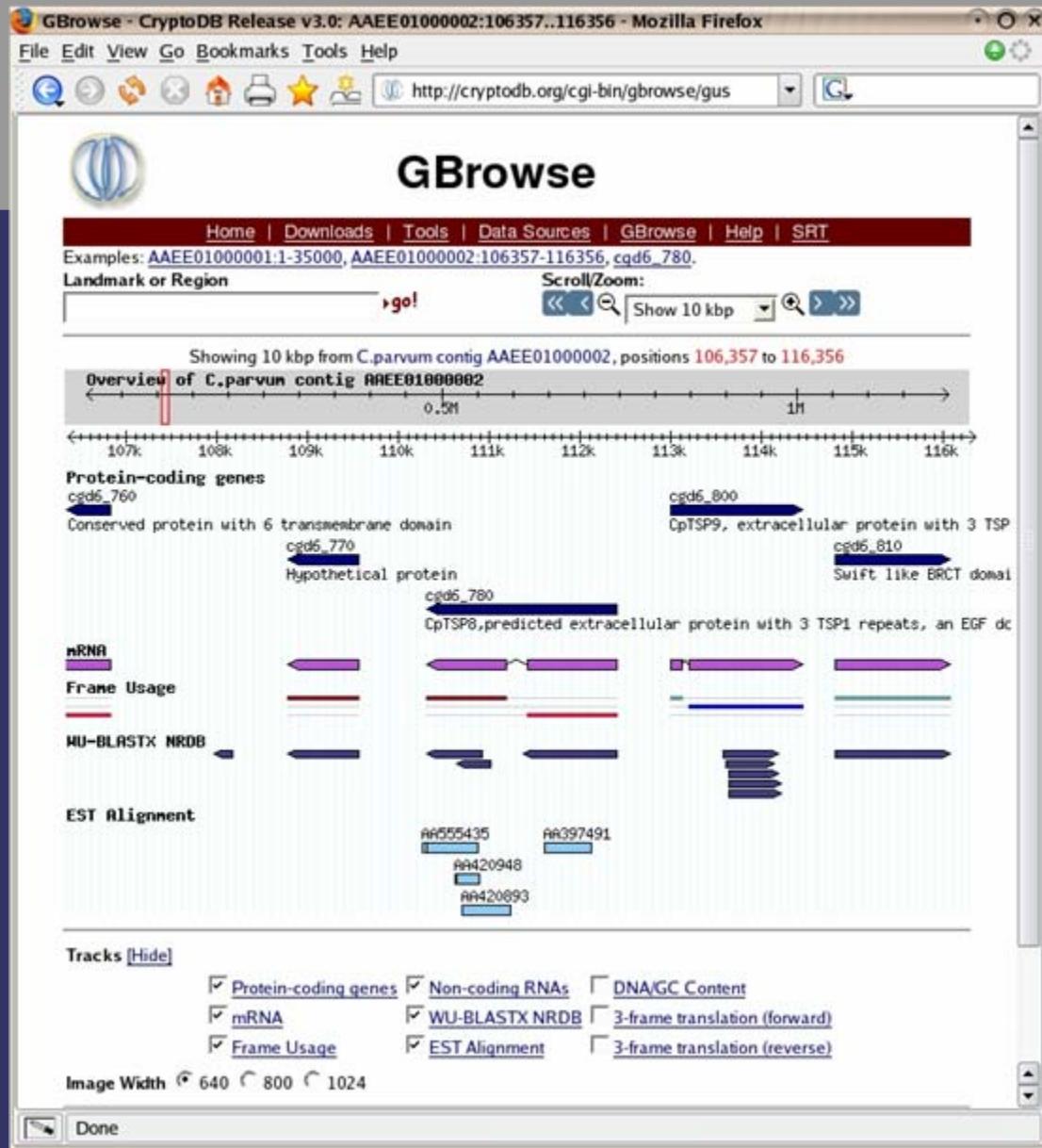
➲ Objectives

- Provide DAS services for all completed sequenced genomes in the ApiComplexa
- Federate genome annotation
- Utilize GBrowse's DAS capability

In other words, a single machine can gather genome annotation information from multiple Web sites, collate the information, and display it to the user in a single view.

ApiComplexa DAS Schematic





Features:

- * Access GUS via DAS adaptor
- * Hierarchical Structure Support
- * Customized Look & Feel
- * Rich Graphical Tooltips

<http://CryptoDB.org>

CryptoDB v3.0 GFF3

CryptoDB GFF3 Snippet – compatible with IOWG GFF3 Convention

```
##gff-version 3
##feature-ontology so.obo
##attribute-ontology gff3_attributes.obo
AAEE01000001    Genbank contig 1      1278458 .      +      .
ID=AAEE01000001;Name=AAEE01000001;molecule_type=dsDNA;Dbxref=taxon:5807,Genbank:AAEE01000001;size=9.11Mb;organism
_name=Cryptosporidium parvum;strain=Iowa type II;translation_table=1;topology=linear;localization=nuclear
AAEE01000001    Genbank gene   6844    9270    .      -      .      ID=gene.63238;Name=cgd7_10
AAEE01000001    Genbank mRNA   6844    9270    .      -      .      ID=mRNA.63239;Parent=gene.63238;Name=cgd7_10;locus=cgd7_10;Dbxref=GI:46229688;description=hypothetical protein%2C
transcript identified by EST
AAEE01000001    Genbank CDS    6844    9270    .      -      .      ID=CDS.63240;Parent=mRNA.63239
AAEE01000001    Genbank exon   6844    9270    .      -      .      ID=exon.63240;Parent=gene.63238
AAEE01000001    Genbank gene   13964   21388   .      +      .      ID=gene.63243;Name=cgd7_30
AAEE01000001    Genbank mRNA   13964   21388   .      +      .      ID=mRNA.63244;Parent=gene.63243;Name=cgd7_30;locus=cgd7_30;Dbxref=GI:46229689;description=large uncharacterized
protein
AAEE01000001    Genbank CDS    13964   21388   .      +      .      ID=CDS.63245;Parent=mRNA.63244
AAEE01000001    Genbank exon   13964   21388   .      +      .      ID=exon.63245;Parent=gene.63243
AAEE01000001    Genbank gene   21516   23990   .      -      .      ID=gene.63247;Name=cgd7_40
AAEE01000001    Genbank mRNA   21516   23990   .      -      .      ID=mRNA.63248;Parent=gene.63247;Name=cgd7_40;locus=cgd7_40;Dbxref=GI:46229844;description=calcium%2Fcalmodulin-
dependent protein kinase with a kinase domain and 2 calmodulin-like EF hands
AAEE01000001    Genbank CDS    21516   23990   .      -      .      ID=CDS.63249;Parent=mRNA.63248
AAEE01000001    Genbank exon   21516   23990   .      -      .      ID=exon.63249;Parent=gene.63247
```

genome_size ?

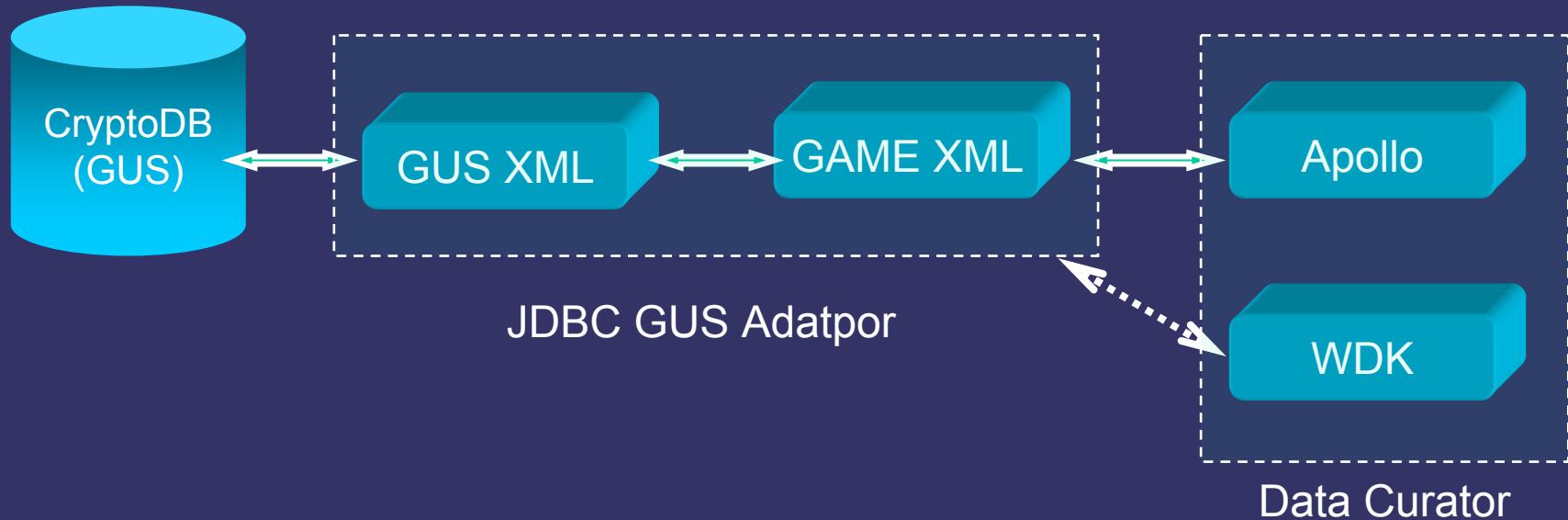
URL Escape

only [a-zA-Z0-9. :^*\$@!+ _?-] are allowed

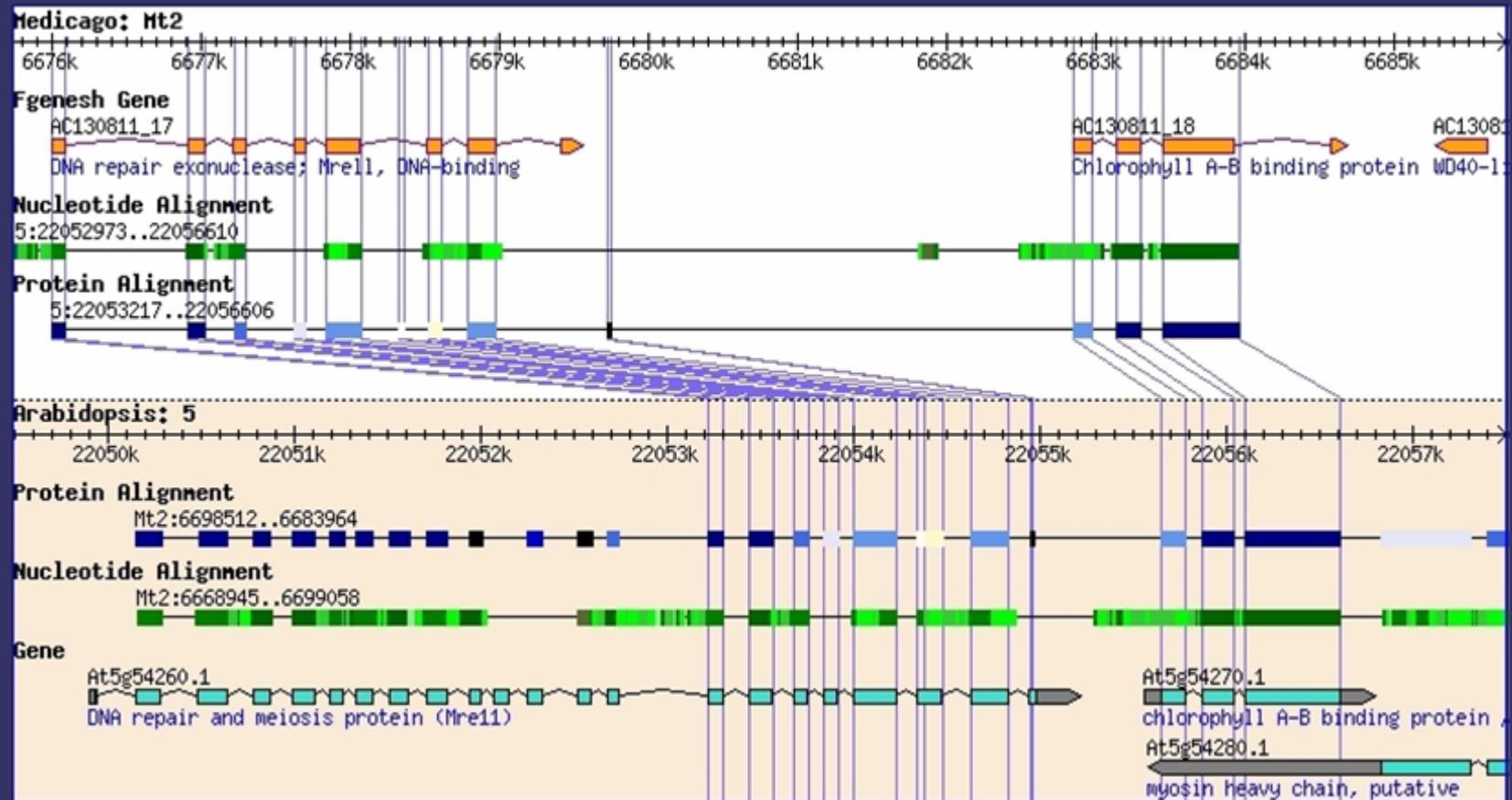
Apollo A Genome Annotation and Curation Tool

↪ Data Flow

- Apollo <-> GAME XML <-> GUS XML <-> GUSDB



SynBrowse: A Synteny Browser for Comparative Sequence Analysis



<http://www.synbrowse.org>

Acknowledgements

University of Georgia (ApiDB/CryptoDB, Kissinger Lab)

Cristina Aurrecoechea
Xin Gao
Congzhou He
Mark Heiges
Conrad Ibanez
Jessica Kissinger
Eileen Kramer
John Miller
Phillippa Rhodes
Ed Robinson
Haiming Wang
Sammy Wang
Yin Xiong

University of Pennsylvania (ApiDB/PlasmoDB/ToxoDB, GUS Team, Roos Lab, Stoeckert Lab)

Amit Bahl Philip Labo
Feng Chen Aaron Mackey
Shailesh Date Jules Milgram
Kobby Essien Deborah Pinney
Steve Fischer David Roos
Bindu Gajria Micheal Saffitz
Thomas Gan Chris Stoeckert
Greg Grant Patricia Whetzel
John Iodice